Introduction to Bayesian Inference: Supplemental Topics

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://hosting.astro.cornell.edu/~loredo/

CASt Summer School — 5–9 June 2023

Supplemental Topics

1. **Estimation and model comparison for binary outcomes; probability & frequency**
2. **Basic inference with normal errors**
3. **Poisson distribution; the on/off problem**
4. **Model uncertainty**
5. **Assigning priors**

Supplemental Topics

1. **Estimation and model comparison for binary outcomes; probability & frequency**
2. **Basic inference with normal errors**
3. **Poisson distribution; the on/off problem**
4. **Model uncertainty**
5. **Assigning priors**

Binary Outcomes: Parameter Estimation

$M =$ Existence of two outcomes, S and F; for each case or trial, the probability for S is α; for F it is $(1-\alpha)$

$H_i =$ Statements about α, the probability for success on the next trial $\rightarrow$ seek $p(\alpha|D,M)$

$D =$ Sequence of results from N observed trials:

FFSSSSFSSSFS ($n = 8$ successes in $N = 12$ trials)

Likelihood:

$$\begin{aligned} p(D|\alpha, M) &= p(\text{failure}|\alpha, M) \times p(\text{failure}|\alpha, M) \times \cdots \\ &= \alpha^n (1-\alpha)^{N-n} \\ &= \mathcal{L}(\alpha) \end{aligned}$$

Prior

Starting with no information about α beyond its definition, use as an "uninformative" prior $p(\alpha|M) = 1$. Justifications:

- Intuition: Don't prefer any α interval to any other of same size
- Bayes's justification: "Ignorance" means that before doing the N trials, we have no preference for how many will be successes:

$$P(n\,\text{success}|M) = \frac{1}{N+1} \qquad \rightarrow \qquad p(\alpha|M) = 1$$

Consider this a *convention*—an assumption added to M to make the problem well posed.

Prior Predictive

$$
\begin{aligned}
p(D|M) &= \int d\alpha\ \alpha^n (1-\alpha)^{N-n} \\
&= B(n+1, N-n+1) = \frac{n!(N-n)!}{(N+1)!}
\end{aligned}
$$

A *Beta integral*, $B(a, b) \equiv \int dx\, x^{a-1}(1-x)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Posterior

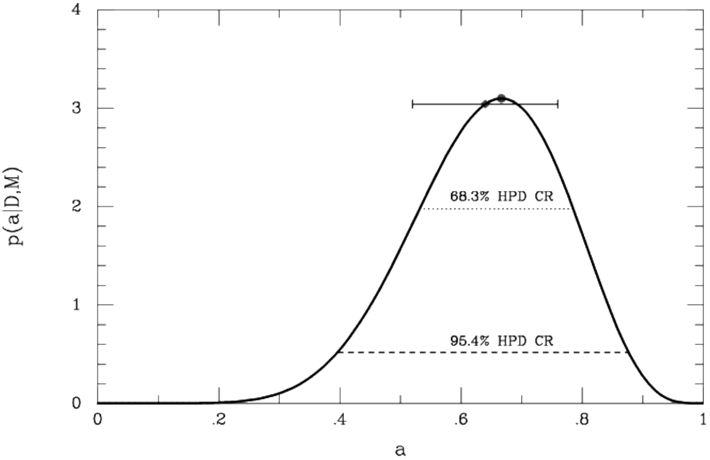
$$p(\alpha|D, M) \quad = \quad \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

A *Beta distribution*. Summaries:

- Best-fit: $\hat{\alpha} = \frac{n}{N} = 2/3$; $\langle\alpha\rangle = \frac{n+1}{N+2} \approx 0.64$
- Uncertainty: $\sigma_\alpha = \sqrt{\frac{(n+1)(N-n+1)}{(N+2)^2(N+3)}} \approx 0.12$
  Find credible regions numerically, or with incomplete beta function

Note that the posterior depends on the data only through n, not the N binary numbers describing the sequence.

n is a (minimal) *sufficient statistic*.



68.3% HPD CR
95.4% HPD CR
p(a|D,M)
a
0
.2
.4
.6
.8
1
0
1
2
3
4

Binary Outcomes: Model Comparison

Equal Probabilities?

M_1: $\alpha = 1/2$
M_2: $\alpha \in [0, 1]$ with flat prior.

Maximum Likelihoods

$$M_1: \qquad p(D|M_1) = \frac{1}{2^N} = 2.44 \times 10^{-4}$$

$$M_2: \quad \mathcal{L}(\hat{\alpha}) = \left(\frac{2}{3}\right)^n \left(\frac{1}{3}\right)^{N-n} = 4.82 \times 10^{-4}$$

$$\frac{p(D|M_1)}{p(D|\hat{\alpha}, M_2)} = 0.51$$

Maximum likelihoods favor M_2 (failures more probable).

Bayes Factor (ratio of model likelihoods)

$$p(D|M_1) = \frac{1}{2^N}; \qquad \text{and} \qquad p(D|M_2) = \frac{n!(N-n)!}{(N+1)!}$$

$$\begin{aligned}\to B_{12} \equiv \frac{p(D|M_1)}{p(D|M_2)} &= \frac{(N+1)!}{n!(N-n)!2^N} \\ &= 1.57\end{aligned}$$

Bayes factor (odds) favors M_1 (equiprobable).

Note that for $n = 6$, $B_{12} = 2.93$; for this small amount of data, we can never be very sure results are equiprobable.

If $n = 0$, $B_{12} \approx 1/315$; if $n = 2$, $B_{12} \approx 1/4.8$; for extreme data, 12 flips *can* be enough to lead us to strongly suspect outcomes have different probabilities.

(Frequentist significance tests can reject null for any sample size.)

Binary Outcomes: Binomial Distribution

Suppose $D = n$ (number of heads in N trials), rather than the actual sequence. What is $p(\alpha|n, M)$?

Likelihood

Let $\mathcal{S}$ = a sequence of flips with n heads.

$$\begin{aligned} p(n|\alpha, M) &= \sum_{\mathcal{S}} \overbrace{p(S|\alpha, M)}^{\alpha^n (1-\alpha)^{N-n}} \; \underbrace{p(n|\mathcal{S}, \alpha, M)}_{[\![\# \text{ successes} = n]\!]} \\ &= \alpha^n (1-\alpha)^{N-n} C_{n,N} \end{aligned}$$

$C_{n,N} = \#$ of sequences of length N with n heads.

$$\rightarrow p(n|\alpha, M) = \frac{N!}{n!(N-n)!} \alpha^n (1-\alpha)^{N-n}$$

The *binomial distribution* for n given α, N.

Posterior

$$p(\alpha|n, M) = \frac{\frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}}{p(n|M)}$$

$$\begin{aligned} p(n|M) &= \frac{N!}{n!(N-n)!}\int d\alpha\; \alpha^n(1-\alpha)^{N-n} \\ &= \frac{1}{N+1} \end{aligned}$$

$$\rightarrow p(\alpha|n, M) = \frac{(N+1)!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n}$$

Same result as when data specified the actual sequence.

Probability & frequency

Frequencies are relevant when modeling repeated trials, or repeated sampling from a population or ensemble.

Frequencies are observables

- When available, can be used to *infer* probabilities for next trial
- When unavailable, can be *predicted*

Bayesian/Frequentist relationships

- Relationships between probability and frequency
- Long-run performance of Bayesian procedures

Probability & frequency in IID settings

Frequency from probability

Bernoulli's law of large numbers: In repeated i.i.d. trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\rm success}}{N_{\rm total}} \to \alpha \quad \text{as} \quad N_{\rm total} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be interpreted as a frequency distribution.

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" $\to$ First use of Bayes's theorem:

Probability for success in next trial of i.i.d. sequence:

$$\text{E}(\alpha) \to \frac{n_{\rm success}}{N_{\rm total}} \quad \text{as} \quad N_{\rm total} \to \infty$$

If $p(x)$ does not change from sample to sample, it may be estimated from a frequency distribution.

The weather forecaster

Joint Frequencies of Actual & Predicted Weather

	Actual		
Prediction	Rain	Sun	
Rain	**1/4**	**1/2**	3/4
Sun	**0**	**1/4**	1/4
	1/4	3/4	

Forecaster is right only 50% of the time

Observer notes a prediction of 'Sun' *every day* would be right 75% of the time, and applies for the forecaster's job

Should the observer get the job?

	Actual	
Prediction	Rain	Sun
Rain	1/4	1/2
Sun	0	1/4

Forecaster: You'll never be in an unpredicted rain

Observer: You'll be in an unpredicted rain 1 day out of 4

Bayesian viewpoint

The value of an inference lies in its usefulness in the individual case

Long run performance is not an adequate criterion for assessing the usefulness of of an inference procedure

When long run performance is deemed important, it needs to be separately evaluated

Supplemental Topics

1. Estimation and model comparison for binary outcomes; probability & frequency
2. **Basic inference with normal errors**
3. Poisson distribution; the on/off problem
4. Model uncertainty
5. Assigning priors

Inference With Normals/Gaussians

Gaussian PDF

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{over } [-\infty, \infty]$$

Common abbreviated notation: $x \sim N(\mu, \sigma^2)$

Parameters

$$\begin{aligned} \mu &= \langle x \rangle \equiv \int dx\, x\, p(x|\mu,\sigma) \\ \sigma^2 &= \langle (x-\mu)^2 \rangle \equiv \int dx\, (x-\mu)^2\, p(x|\mu,\sigma) \end{aligned}$$

Gauss's Observation: Sufficiency

Suppose our data consist of N measurements, $d_i = \mu + \epsilon_i$. Suppose the noise contributions are independent, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

$$
\begin{aligned}
p(D|\mu, \sigma, M) &= \prod_i p(d_i|\mu, \sigma, M) \\
&= \prod_i p(\epsilon_i = d_i - \mu|\mu, \sigma, M) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(d_i - \mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{\sigma^N (2\pi)^{N/2}} e^{-Q(\mu)/2\sigma^2}
\end{aligned}
$$

Find dependence of Q on μ by completing the square:

$$
\begin{aligned}
Q &= \sum_i (d_i - \mu)^2 \qquad\qquad [\text{Note: } Q/\sigma^2 = \chi^2(\mu)] \\
&= \sum_i d_i^2 + \sum_i \mu^2 - 2\sum_i d_i \mu \\
&= \left(\sum_i d_i^2\right) + N\mu^2 - 2N\mu\overline{d} \qquad \text{where } \overline{d} \equiv \frac{1}{N}\sum_i d_i \\
&= N(\mu - \overline{d})^2 + \left(\sum_i d_i^2\right) - N\overline{d}^2 \\
&= N(\mu - \overline{d})^2 + Nr^2 \quad \text{where } r^2 \equiv \frac{1}{N}\sum_i (d_i - \overline{d})^2
\end{aligned}
$$

Likelihood depends on $\{d_i\}$ only through $\overline{d}$ and r:

$$\mathcal{L}(\mu, \sigma) = \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)$$

The sample mean and variance are *sufficient statistics*.

This is a miraculous compression of information—the normal dist'n is highly *abnormal* in this respect!

Estimating a Normal Mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is known $\rightarrow I = (\sigma, M)$.

Parameter space: μ; seek $p(\mu|D, \sigma, M)$

Likelihood

$$
\begin{aligned}
p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)
\end{aligned}
$$

"Uninformative" prior

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.
This prior is *improper* unless bounded.

Prior predictive/normalization

$$\begin{aligned} p(D|\sigma, M) &= \int d\mu\; C \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &= C(\sigma/\sqrt{N})\sqrt{2\pi} \end{aligned}$$

... minus a tiny bit from tails, using a proper prior.

Posterior

$$p(\mu|D, \sigma, M) = \frac{1}{(\sigma/\sqrt{N})\sqrt{2\pi}} \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)$$

Posterior is $N(\overline{d}, w^2)$, with standard deviation $w = \sigma/\sqrt{N}$.

68.3% HPD credible region for μ is $\overline{d} \pm \sigma/\sqrt{N}$.

Note that C drops out $\rightarrow$ limit of infinite prior range is well behaved.

Informative Conjugate Prior

Use a normal prior, $\mu \sim N(\mu_0, w_0^2)$.

Conjugate because the posterior turns out also to be normal.

Posterior

Normal $N(\tilde{\mu}, \tilde{w}^2)$, but mean, std. deviation *"shrink"* towards prior.

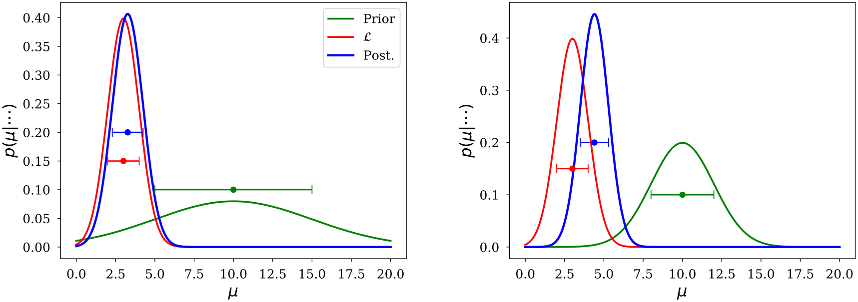
Define $B = \frac{w^2}{w^2 + w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large. Then

$$\begin{aligned} \widetilde{\mu} &= \overline{d} + B \cdot (\mu_0 - \overline{d}) \\ \widetilde{w} &= w \cdot \sqrt{1 - B} \end{aligned}$$

"Principle of stable estimation" — The prior affects estimates only when data are not informative relative to prior.

Conjugate normal examples:

- Data have $\overline{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$



Estimating a Normal Mean: Unknown σ

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is *unknown*

Parameter space: (μ, σ); seek $p(\mu|D, M)$

Likelihood

$$\begin{aligned} p(D|\mu, \sigma, M) &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left(-\frac{Nr^2}{2\sigma^2}\right) \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right) \\ &\propto \frac{1}{\sigma^N} e^{-Q/2\sigma^2} \end{aligned}$$

$$\text{where} \quad Q = N\left[r^2 + (\mu - \overline{d})^2\right]$$

Uninformative Priors

Assume priors for μ and σ are independent.

Translation invariance $\Rightarrow p(\mu) \propto C$, a constant.

Scale invariance $\Rightarrow p(\sigma) \propto 1/\sigma$ (flat in $\log \sigma$).

Joint Posterior for μ, σ

$$p(\mu, \sigma | D, M) \propto \frac{1}{\sigma^{N+1}} e^{-Q(\mu)/2\sigma^2}$$

Marginal Posterior

$$p(\mu|D,M) \propto \int d\sigma \, \frac{1}{\sigma^{N+1}} e^{-Q/2\sigma^2}$$

Let $\tau = \frac{Q}{2\sigma^2}$ so $\sigma = \sqrt{\frac{Q}{2\tau}}$ and $|d\sigma| = \tau^{-3/2}\sqrt{\frac{Q}{2}}\, d\tau$

$$\begin{aligned} \Rightarrow p(\mu|D,M) &\propto 2^{N/2} Q^{-N/2} \int d\tau \, \tau^{\frac{N}{2}-1} e^{-\tau} \\ &\propto Q^{-N/2} \end{aligned}$$

Write $Q = Nr^2\left[1+\left(\frac{\mu-\overline{d}}{r}\right)^2\right]$ and normalize:

$$p(\mu|D,M) = \frac{\left(\frac{N}{2}-1\right)!}{\left(\frac{N}{2}-\frac{3}{2}\right)!\sqrt{\pi}}\frac{1}{r}\left[1+\frac{1}{N}\left(\frac{\mu-\overline{d}}{r/\sqrt{N}}\right)^2\right]^{-N/2}$$

"Student's t distribution," with $t = \frac{(\mu-\overline{d})}{r/\sqrt{N}}$

A "bell curve," but with power-law tails

Large N:

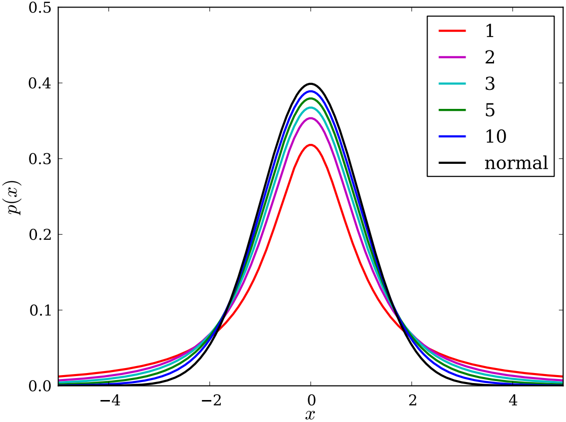
$$p(\mu|D,M) \sim e^{-N(\mu-\overline{d})^2/2r^2}$$

This is the rigorous way to "adjust σ so $\chi^2/\mathrm{dof} = 1$."

It doesn't just plug in a best σ; it slightly broadens posterior to account for σ uncertainty.

Student t examples:

- $p(x) \propto \frac{1}{\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}}$
- Location $= 0$, scale $= 1$
- Degrees of freedom $= \{1, 2, 3, 5, 10, \infty\}$



Gaussian Background Subtraction

Measure background rate $b = \hat{b} \pm \sigma_b$ with source off.

Measure total rate $r = \hat{r} \pm \sigma_r$ with source on.

Infer signal source strength s, where $r = s + b$.

With flat priors,

$$p(s, b|D, M) \propto \exp\left[-\frac{(b - \hat{b})^2}{2\sigma_b^2}\right] \times \exp\left[-\frac{(s + b - \hat{r})^2}{2\sigma_r^2}\right]$$

Marginalize b to summarize the results for s (complete the square to isolate b dependence; then do a simple Gaussian integral over b):

$$p(s|D, M) \propto \exp\left[-\frac{(s-\hat{s})^2}{2\sigma_s^2}\right] \qquad \begin{array}{l} \hat{s} = \hat{r} - \hat{b} \\ \sigma_s^2 = \sigma_r^2 + \sigma_b^2 \end{array}$$

$\Rightarrow$ Background *subtraction* is a special case of background *marginalization*; i.e., marginalization "told us" to subtract a background estimate.

Recall the standard derivation of background uncertainty via "propagation of errors" based on Taylor expansion (statistician's *Delta-method*).

Marginalization provides a generalization of error propagation—without approximation!

Supplemental Topics

1. Estimation and model comparison for binary outcomes; probability & frequency
2. Basic inference with normal errors
3. **Poisson distribution; the on/off problem**
4. Model uncertainty
5. Assigning priors

Poisson Dist'n: Infer a Rate from Counts

Problem:

Observe n counts in T; infer rate, r

Likelihood

$$\mathcal{L}(r) \equiv p(n|r, M) = p(n|r, M) = \frac{(rT)^n}{n!} e^{-rT}$$

Prior

Two simple standard choices (or conjugate gamma dist'n):

- r known to be nonzero; it is a scale parameter:

$$p(r|M) = \frac{1}{\ln(r_u/r_l)} \frac{1}{r}$$

- r may vanish; require $p(n|M) \sim$ Const:

$$p(r|M) = \frac{1}{r_u}$$

Prior predictive

$$
\begin{aligned}
p(n|M) &= \frac{1}{r_u}\frac{1}{n!}\int_0^{r_u} dr (rT)^n e^{-rT} \\
&= \frac{1}{r_u T}\frac{1}{n!}\int_0^{r_u T} d(rT)(rT)^n e^{-rT} \\
&\approx \frac{1}{r_u T} \quad \text{for} \quad r_u \gg \frac{n}{T}
\end{aligned}
$$

Posterior

A gamma distribution:

$$
p(r|n, M) = \frac{T(rT)^n}{n!} e^{-rT}
$$

Gamma Distributions

A 2-parameter family of distributions over nonnegative x, with shape parameter α and scale parameter s:

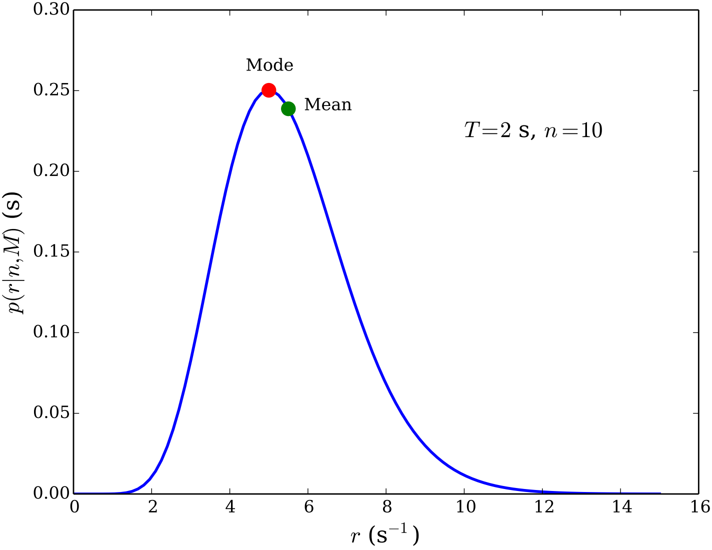
$$p_\Gamma(x|\alpha, s) \;=\; \frac{1}{s\Gamma(\alpha)}\left(\frac{x}{s}\right)^{\alpha-1} e^{-x/s}$$

Moments:

$$\mathrm{E}(x) = s\alpha \qquad \mathrm{Var}(x) = s^2\alpha$$

Our posterior corresponds to $\alpha = n + 1$, $s = 1/T$.

- Mode $\hat{r} = \frac{n}{T}$; mean $\langle r \rangle = \frac{n+1}{T}$ (shift down 1 with $1/r$ prior)
- Std. dev'n $\sigma_r = \frac{\sqrt{n+1}}{T}$; credible regions found by integrating (can use incomplete gamma function)



Mode
Mean
T=2 s, n=10
0.00
0.05
0.10
0.15
0.20
0.25
0.30
$p(r|n,M)$ (s)
0
2
4
6
8
10
12
14
16
r (s^{-1})

Conjugate prior

Note that a gamma distribution prior is the conjugate prior for the Poisson sampling distribution:

$$\begin{aligned} p(r|n, M') &\propto \mathrm{Gamma}(r|\alpha, s) \times \mathrm{Pois}(n|rT) \\ &\propto r^{\alpha-1} e^{-r/s} \times r^n e^{-rT} \\ &\propto r^{\alpha+n-1} \exp[-r(T + 1/s)] \end{aligned}$$

For $\alpha = 1$, $s \to \infty$, the gamma prior becomes an "uninformative" flat prior; posterior is proper even for $n = 0$

Useful conventions

- Use a flat prior for a rate that may be zero
- Use a log-flat prior ($\propto 1/r$) for a nonzero scale parameter
- Use proper (normalized, bounded) priors
- Plot posterior with abscissa that makes prior flat (use $\log r$ abscissa for scale parameter case)

Infer a Signal in a Known Background

Problem:

As before, but $r = s + b$ with b known; infer s

$$p(s|n, b, M) = C \frac{T\left[(s+b)T\right]^n}{n!} e^{-(s+b)T}$$

$$\begin{aligned} C^{-1} &= \frac{e^{-bT}}{n!} \int_0^\infty d(sT)\,(s+b)^n T^n e^{-sT} \\ &= \sum_{i=0}^{n} \frac{(bT)^i}{i!} e^{-bT} \end{aligned}$$

A sum of Poisson probabilities for background events; it can be evaluated using the incomplete gamma function. (Helene 1983)

The On/Off Problem

Basic problem

- Look off-source; unknown background rate b
  Count $N_{\rm off}$ photons in interval $T_{\rm off}$
- Look on-source; rate is $r = s + b$ with unknown signal s
  Count $N_{\rm on}$ photons in interval $T_{\rm on}$
- Infer s

Conventional solution

$$\begin{aligned}
\hat{b} = N_{\rm off}/T_{\rm off}; &\quad \sigma_b = \sqrt{N_{\rm off}}/T_{\rm off} \\
\hat{r} = N_{\rm on}/T_{\rm on}; &\quad \sigma_r = \sqrt{N_{\rm on}}/T_{\rm on} \\
\hat{s} = \hat{r} - \hat{b}; &\quad \sigma_s = \sqrt{\sigma_r^2 + \sigma_b^2}
\end{aligned}$$

But $\hat{s}$ can be *negative*!

Bayesian Solution to On/Off Problem

First consider off-source data; use it to estimate b:

$$p(b|N_{\rm off}, I_{\rm off}) = \frac{T_{\rm off}(bT_{\rm off})^{N_{\rm off}} e^{-bT_{\rm off}}}{N_{\rm off}!}$$

Use this as a prior for b to analyze on-source data. For on-source analysis $I_{\rm all} = (I_{\rm on}, N_{\rm off}, I_{\rm off})$:

$$p(s, b|N_{\rm on}) \quad \propto \quad p(s)p(b)[(s+b)T_{\rm on}]^{N_{\rm on}} e^{-(s+b)T_{\rm on}} \qquad || \; I_{\rm all}$$

$p(s|I_{\rm all})$ is flat, but $p(b|I_{\rm all}) = p(b|N_{\rm off}, I_{\rm off})$, so

$$p(s, b|N_{\rm on}, I_{\rm all}) \quad \propto \quad (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}$$

Now marginalize over b;

$$
\begin{aligned}
p(s|N_{\rm on}, I_{\rm all}) &= \int db\; p(s, b \mid N_{\rm on}, I_{\rm all}) \\
&\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}
\end{aligned}
$$

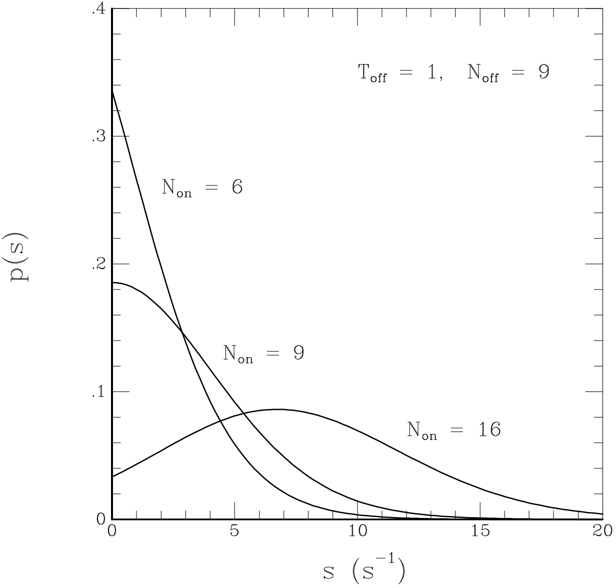
Expand $(s+b)^{N_{\rm on}}$ and do the resulting Γ integrals:

$$
\begin{aligned}
p(s|N_{\rm on}, I_{\rm all}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\
C_i &\propto \left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on}+N_{\rm off}-i)!}{(N_{\rm on}-i)!}
\end{aligned}
$$

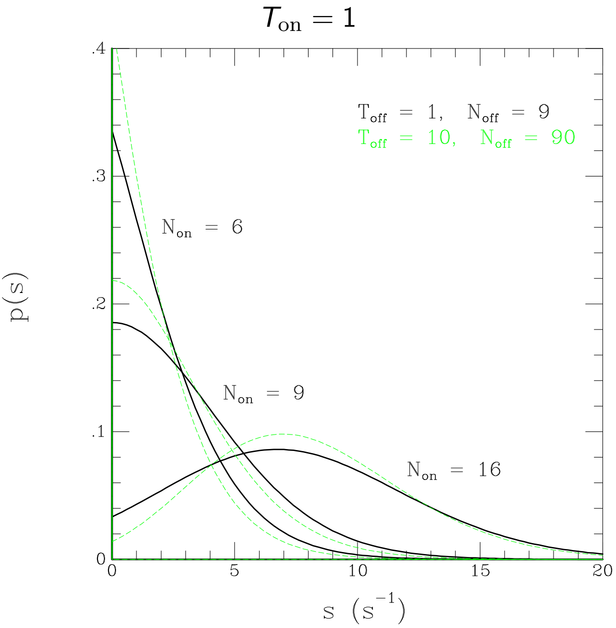
Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

Example On/Off Posteriors—Short Integrations

$T_{\rm on} = 1$



Example On/Off Posteriors—Long Background Integrations



Second Solution of the On/Off Problem

Consider all the data at once; the likelihood is a product of Poisson distributions for the on- and off-source counts:

$$
\begin{aligned}
\mathcal{L}(s,b) &\equiv p(N_{\rm on}, N_{\rm off}|s,b,I) \\
&\propto [(s+b)T_{\rm on}]^{N_{\rm on}} e^{-(s+b)T_{\rm on}} \times (bT_{\rm off})^{N_{\rm off}} e^{-bT_{\rm off}}
\end{aligned}
$$

Take joint prior to be flat; find the joint posterior and marginalize over b;

$$
\begin{aligned}
p(s|N_{\rm on}, I_{\rm on}) &= \int db\ p(s,b|I)\,\mathcal{L}(s,b) \\
&\propto \int db\ (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}
\end{aligned}
$$

$\rightarrow$ same result as before.

A profound consistency

We solved the on/off problem in multiple ways, finding the same final results.

This reflects something fundamental about Bayesian inference.

R. T. Cox proposed two necessary conditions for a quantification of uncertainty:

- It should duplicate deductive logic when there is no uncertainty
- Different decompositions of arguments should produce the same final quantifications (internal consistency)

Surprising result: These conditions (formalized) are *sufficient*; they require uncertainty quantification to follow the rules of Bayesian probability theory. E. T. Jaynes and others refined and simplified Cox's analysis.

Multibin On/Off

The more typical on/off scenario:

Data = spectrum or image with counts in many bins

Model M gives signal rate $s_k(\theta)$ in bin k, parameters θ

To infer θ, we need the likelihood:

$$\mathcal{L}(\theta) = \prod_k p(N_{\mathrm{on}\,k}, N_{\mathrm{off}\,k} | s_k(\theta), M)$$

For each k, we have an on/off problem as before, only we just need the marginal likelihood for s_k (not the posterior). The same C_i coefficients arise.

XSPEC and CIAO/Sherpa provide this as an option (maybe not the latest Sherpa)

van Dyk$^+$(2001) does the same thing via data augmentation (Monte Carlo)

Supplemental Topics

1. Estimation and model comparison for binary outcomes; probability & frequency
2. Basic inference with normal errors
3. Poisson distribution; the on/off problem
4. **Model uncertainty**
5. Assigning priors

Model comparison

Problem statement

$\mathcal{C} = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$\begin{aligned} p(M_i|D,\mathcal{C}) &= p(M_i|\mathcal{C})\frac{p(D|M_i,\mathcal{C})}{p(D|\mathcal{C})} \\ &\propto p(M_i|\mathcal{C})\mathcal{L}(M_i) \end{aligned}$$

$$\mathcal{L}(M_i) \equiv p(D|M_i) = \int d\theta_i \, p(\theta_i|M_i)p(D|\theta_i, M_i)$$

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = *Marginal likelihood* = Average likelihood = Global likelihood = (Weight of) Evidence for model

Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the *odds* favoring one over the other:

$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D,\mathcal{C})}{p(M_j|D,\mathcal{C})} \\ &= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})} \end{aligned}$$

The data-dependent part is called the *Bayes factor*:

$$B_{ij} \equiv \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods for *composite hypotheses*

An automatic Ockham's razor

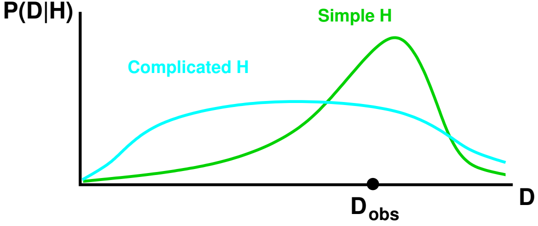
Consider *nested models*:

- Simpler model M_1 with parameters θ_1
- "Larger" rival M_2 with parameters $\theta_2 = (\theta_1, \eta)$

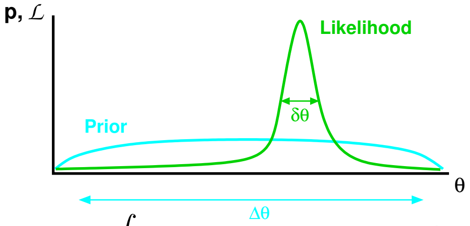
$\Rightarrow \mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$
But what about $p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i)$?

Prior predictive distributions

Normalization implies *there must be data that favor M_1*:



The Ockham factor

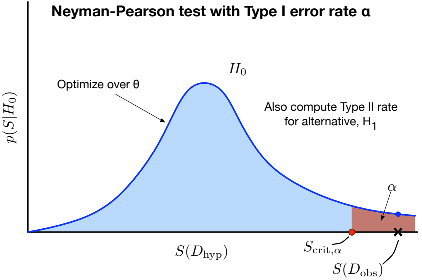


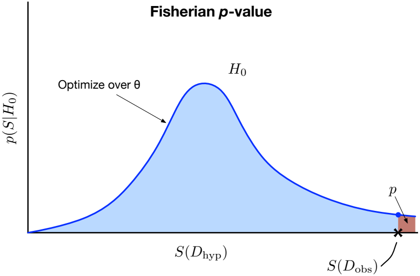
$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Ockham Factor}
\end{aligned}
$$

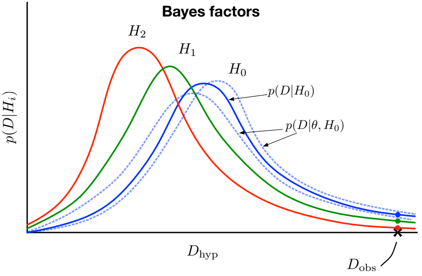
Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for "wasted" volume of parameter space

Quantifies intuition that models shouldn't require fine-tuning







- NP & Fisher give H_0 a special role
- NP & Fisher optimize over θ, integrate over $D_{\rm hyp}$
- Bayes considers rival H_i symmetrically
- Bayes integrates over θ, uses only $D_{\rm obs}$

Bayes factors can only compare rival models;
they don't measure "goodness-of-fit"

Posterior predictive p-values are a BDA alternative
for measuring "suprisingness" of data for model checking;
they integrate over both data and parameter spaces

See "p-value note" online for 2016 CASt summer; or 2018 Sagan workshop slides

Model averaging

Problem statement

$I = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models
Models all share a set of "interesting" parameters, ϕ
Each has different set of nuisance parameters η_i (or different prior info about them)
H_i = statements about ϕ

Model averaging

Calculate posterior PDF for ϕ:

$$\begin{aligned} p(\phi|D,\mathcal{C}) &= \sum_i p(M_i|D,\mathcal{C})\, p(\phi|D,M_i) \\ &\propto \sum_i \mathcal{L}(M_i) \int d\eta_i\, p(\phi,\eta_i|D,M_i) \end{aligned}$$

The model choice is a (discrete) nuisance parameter here

Useful for handling *systematic error* in estimation & prediction

Supplemental Topics

1. Estimation and model comparison for binary outcomes; probability & frequency

2. Basic inference with normal errors

3. Poisson distribution; the on/off problem

4. Model uncertainty

5. **Assigning priors**

Well-Posed Problems

The rules (BT, LTP, . . .) express desired probabilities in terms of other probabilities

To get a numerical value *out*, at some point we have to put numerical values *in*

Direct probabilities are probabilities with numerical values determined directly by premises (via modeling assumptions, symmetry arguments, previous calculations, desperate presumption . . .)

An inference problem is *well posed* only if all the needed probabilities are assignable based on the context. We may need to add new assumptions as we see what needs to be assigned. We may not be entirely comfortable with what we need to assume! (Remember Euclid's fifth postulate!)

Should explore how results depend on uncomfortable assumptions ("robustness")

Contextual/prior/background information

Bayes's theorem moves the data and hypothesis propositions wrt the solidus:

$$P(H_i|D_{\text{obs}}, I) = P(H_i|I)\,\frac{P(D_{\text{obs}}|H_i, I)}{P(D_{\text{obs}}|I)}$$

It lets us *change the premises*

"Prior information" or "background information" or "context" = information that is **always** a premise (for the current calculation)

Notation: $P(\cdot|\cdot, I)$ or $P(\cdot|\cdot, \mathcal{C})$ or $P(\cdot|\cdot, M)$ or ...

The context can be a notational nuisance! "Skilling conditional":

$$P(H_i|D_{\text{obs}}) = P(H_i)\,\frac{P(D_{\text{obs}}|H_i)}{P(D_{\text{obs}})} \qquad \| \, \mathcal{C}$$

Essential contextual information

We can only be uncertain about A if there are alternatives; what they are will bear on our uncertainty. *We must explicitly specify relevant alternatives.*

Hypothesis space: The set of alternative hypotheses of interest (and auxiliary hypotheses needed to predict the data)

Data/sample space: The set of possible data we may have predicted before learning of the observed data

Predictive model: Information specifying the likelihood function (e.g., the conditional predictive dist'n/sampling dist'n)

Other prior information: Any further information available or necessary to assume to make the problem *well posed*

Bayesian literature often uses **model** to refer to *all* of the contextual information used to study a particular dataset and predictive model

Directly assigned sampling distributions

Some examples of reasoning leading to sampling distributions:

- Binomial distribution:
  - Ansatz: Probability for a Bernoulli trial, α
  - LTP $\Rightarrow$ binomial for n successes in N trials
- Poisson distribution:
  - Ansatz: $P(\text{event in } dt|\lambda) \propto \lambda dt$;
    probabilities for events in disjoint intervals independent
  - Product & sum rules $\Rightarrow$ Poisson for n in T
- Gaussian distribution:
  - CLT: Probability theory for sum of many quantities with independent, finite-variance PDFs
  - Sufficiency (Gauss): Seek distribution with sample mean as sufficient statistic (also sample variance)
  - Asymptotic limits: large n Binomial, Poisson
  - Others: Herschel's invariance argument (2-D), maximum entropy...

Assigning priors

Sources of prior information

- Analysis of previous experimental/observational data (but begs the question of what prior to use for the first such analysis)
- *Subjective priors*: Elicit a prior from an expert in the problem domain, e.g., via ranges, moments, quantiles, histograms
- *Population priors*: When it's meaningful to pool observations, we potentially can *learn* a shared prior—hierarchical/multilevel modeling does this

"Non-informative" priors

- Seek a prior that in some sense (TBD!) expresses a lack of information prior to considering the data
- No universal solution—this notion must be problem-specific, e.g., exploiting symmetries

Priors derived from the likelihood function

Location/scale problems often admit a transformation group argument identifing good "noninformative" priors.
In other settings we need a more general approach to formal assignment of priors that express "ignorance" in some sense.

There is no universal consensus on how to do this (yet? ever?)

A common underlying idea: The same $\mathcal{C}$ appears in the prior, $p(\theta|\mathcal{C})$, and the likelihood, $p(D|\theta, \mathcal{C})$—the prior "knows" about the likelihood function, although it doesn't know what data values will be plugged into it

Jeffreys priors: Uses Fisher information to define a (parameter-dependent) scale defining a prior; parameterization invariant, but strange behavior in many dimensions

Reference priors: Uses information theory to define a prior that (asymptotically) has the least effect on the posterior; complicated algorithm; gives good frequentist behavior to Bayesian inferences

"Objective" priors

Heuristic motivation:

- Dimensionally, $\pi(\theta) \propto 1/(\theta \text{ scale})$
- Use the likelihood function to determine a (relative) scale at each θ, say, $s(\theta) \rightarrow \pi(\theta) \propto 1/s(\theta)$
- Seek a scale definition that is consistent WRT reparameterization

Such a prior essentially specifies a way to slice-and-dice the θ axis so assigning equal probability to intervals reflects intrinsic scales in the problem, and is consistent WRT reparameterization

Jeffreys priors

Heuristic motivation:

- If we have data D, a natural inverse scale at θ, from the likelihood function, is the **square root** of the *observed Fisher information* (recall Laplace approximation):

$$I_D(\theta) \equiv -\frac{\mathrm{d}^2 \log \mathcal{L}_D(\theta)}{\mathrm{d}\theta^2}$$

- For a prior, we don't know D; for each θ, average over D predicted by the sampling distribution; this defines the *(expected) Fisher information*:

$$I(\theta) \equiv -\mathbb{E}_{D|\theta}\left[\frac{\mathrm{d}^2 \log \mathcal{L}_D(\theta)}{\mathrm{d}\theta^2}\right]$$

- Invariance: Can show for $\phi = \Phi(\theta)$, and $\theta = \Theta(\phi)$:

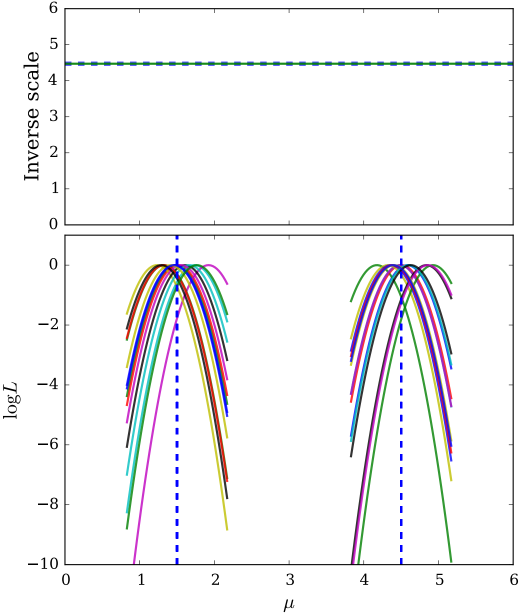
$$I(\phi) = I(\theta)\left(\frac{\mathrm{d}\Theta}{\mathrm{d}\phi}\right)^2$$

Jeffreys' prior:

$$\pi(\theta) \propto [I(\theta)]^{1/2}$$

- Note the proportionality—the prior scale depends on how much the likelihood function scale *changes* vs. θ
- Puts more weight in regions of parameter space where the data are expected to be more informative
- Parameterization invariant, due to use of derivatives and vanishing expectation of the *score function* $S_D(\theta) = \dfrac{\mathrm{d} \log \mathcal{L}_D(\theta)}{\mathrm{d}\theta}$
- Typically improper when parameter space is non-compact
- Improves *frequentist* performance of posterior intervals w.r.t. intervals based on flat priors
- Only considered sound for a single parameter (or considering a single parameter at a time in some multiparameter problems)

Jeffreys prior for normal mean



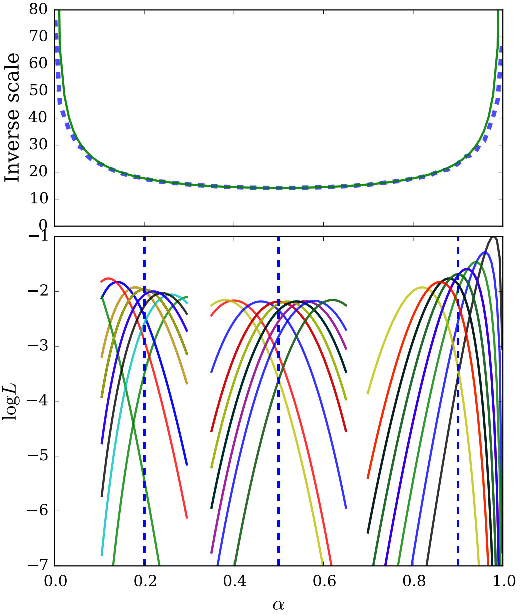
$N = 20$ samples from normals with $\sigma = 1$

Likelihood width is independent of $\mu \Rightarrow$

$$\pi(\mu) = \text{Const}$$

Another justification of the uniform prior
Prior is improper without prior limits on the range

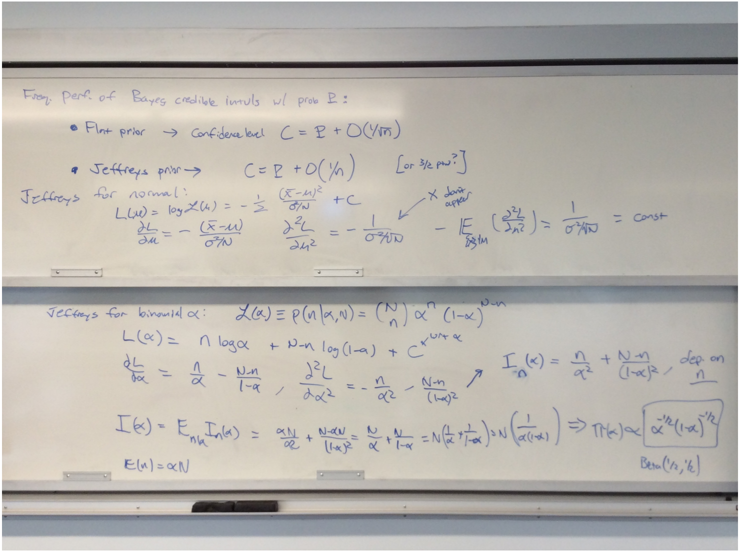
Jeffreys prior for binomial probability



Binomial success counts n from $N = 50$ trials

$$\pi(\mu) = \frac{1}{\pi \alpha^{1/2}(1-\alpha)^{1/2}} = \mathrm{Beta}(1/2, 1/2)$$

Analytical calculations for normal, binomial



Limitations of the Jeffreys prior

- Only considered sound for a single parameter (or considering a single parameter at a time in some multiparameter problems)
  E.g., for $\mathrm{Norm}(\mu, \sigma)$, the Jeffreys prior is $\propto 1/\sigma^2$, *not* the product of separate Jeffreys μ, σ priors
- Only applicable to continuous spaces

$\rightarrow$ Seek more general notions of “objective” or “uninformative” that reproduce good things about the Jeffreys prior

Uncertainty, information, and entropy

Other rules for assigning "non-informative" priors rely on a more formal measure of the *information content* (or its complement, amount of *uncertainty*) in a probability distribution

Intuitively appealing metric-based measures, like standard deviation or interval size, are not general enough; e.g., they don't apply to categorical distributions

Desiderata for an *uncertainty functional* $\mathcal{S}_N[\vec{p}]$—a map from a PMF $\vec{p} = (p_1, p_2, \ldots, p_N)$ to a single scalar quantifying its uncertainty (treat PDFs later):

- $\mathcal{S}_N[\vec{p}]$ should be continuous w.r.t. the p_is
- *Uncertainty grows with multiplicity*: When the p_i are all equal, $s(N) = \mathcal{S}_N[\vec{p}]$ should grow monotonically with N
- *Additivity over subgroups*

Information Gain as Entropy Change

Entropy and uncertainty

Shannon entropy = a scalar measure of the degree of uncertainty expressed by a probability distribution

$$\begin{aligned} \mathcal{S} &= \sum_i p_i \log \frac{1}{p_i} \qquad \text{“Average surprisal”} \\ &= -\sum_i p_i \log p_i \end{aligned}$$

Information gain

Information gain upon learning D = decrease in uncertainty:

$$\begin{aligned} \mathcal{I}(D) &= \mathcal{S}[\{p(H_i)\}] - \mathcal{S}[\{p(H_i|D)\}] \\ &= \sum_i p(H_i|D) \log p(H_i|D) - \sum_i p(H_i) \log p(H_i) \end{aligned}$$

A 'Bit' About Entropy

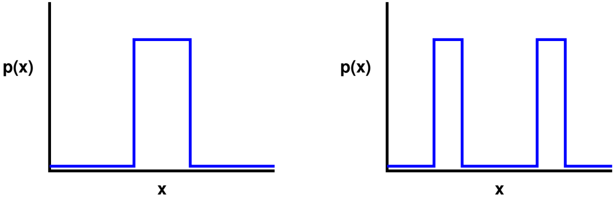
Entropy of a Gaussian

$$p(x) \propto e^{-(x-\mu)^2/2\sigma^2} \quad \rightarrow \quad \mathcal{S} \propto \log(\sigma)$$

$$p(\vec{x}) \propto \exp\left[-\tfrac{1}{2}\vec{x}\cdot \mathrm{V}^{-1}\cdot\vec{x}\right] \quad \rightarrow \quad \mathcal{S} \propto \log(\det \mathrm{V})$$

$\rightarrow$ Asymptotically like log Fisher matrix

A log-measure of "volume" or "spread," not range



These distributions have the same entropy/amount of information.

Expected information gain

When the data are yet to be considered, the *expected* information gain averages over D; straightforward use of the product rule/Bayes's theorem gives:

$$\begin{aligned}
\mathbb{E}\mathcal{I} &= \int dD\, p(D)\, \mathcal{I}(D) \\
&= \int dD\, p(D)\, \sum_i p(H_i|D) \log\left[\frac{p(H_i|D)}{p(H_i)}\right]
\end{aligned}$$

For a continuous hypothesis space labeled by parameter(s) θ,

$$\mathbb{E}\mathcal{I} = \int dD\, p(D)\, \int d\theta p(\theta|D) \log\left[\frac{p(\theta|D)}{p(\theta)}\right]$$

This is the expectation value of the (data-dependent) *Kullback-Leibler divergence* between the prior and posterior:

$$\mathcal{D} \equiv \int d\theta\, p(\theta|D) \log\left[\frac{p(\theta|D)}{p(\theta)}\right]$$

Reference priors

Bernardo (later joined by Berger & Sun) advocates *reference priors*, priors chosen to maximize the KLD between prior and posterior, as an "objective" expression of the idea of a "non-informative" prior: reference priors let the data most strongly dominate the prior (on average)

- Rigorous definition invokes asymptotics and delicate handling of non-compact parameter spaces to make sure posteriors are proper
- For 1-D problems, the reference prior is the Jeffreys prior
- In higher dimensions, the reference prior is *not* the Jeffreys prior; it behaves better
- The construction in higher dimensions is complicated and depends on separating interesting vs. nuisance parameters (but see Berger, Bernardo & Sun 2015, "Overall objective priors")
- Reference priors are typically improper on non-compact spaces
- They give Bayesian inferences good frequentist properties
- A constructive numerical algorithm exists (from particle physicists)